Building a lexicon of French deverbal nouns from a semantically annotated corpus

Antonio Balvet, Lucie Barque, Rafael Marin

Building a lexicon of French deverbal nouns from a semantically annotated corpus

Antonio Balvet, Lucie Barque, Rafael Marín
Univ. Lille Nord de France, F-59000 Lille, France
CNRS, UMR 8163, F-59653
Villeneuve d'Ascq, France
E-mail: name.surname@univ-lille3.fr

Abstract

The ongoing project Nomage aims at describing the aspectual properties of deverbal nouns in an empirical way. It is centered on the development of two resources: a semantically annotated corpus of deverbal nouns, and an electronic lexicon. They are both presented in this paper, and emphasize how the semantic annotations of the corpus allow the lexicographic description of deverbal nouns to be validated, in particular their polysemy.

1. Introduction

From the work of (Lees, 1960), through (Chomsky, 1970) and (Grimshaw, 1990), to more recent studies, nominalizations have occupied a central place in grammatical analysis, with a focus on morphological and syntactic aspects. More recently, researchers have begun to address a specific issue often neglected before, i.e. the semantics of nominalizations, and its implications for Natural Language Processing applications such as electronic ontologies or Information Retrieval.

We focus on precisely this issue in the research project NOMAGE, funded by the French National Research Agency (ANR-07-JCJC-0085-01). This ongoing project aims at describing the aspectual properties of deverbal nouns in an empirical way. It is centered on the development of two resources: a semantically annotated corpus of deverbal nouns, and an electronic lexicon.

In this paper, we present the Nomage corpus and the annotations we make on deverbal nouns (section 2). We then show how we build our lexicon with the semantically annotated corpus and illustrate the kind of generalizations we can make from such data (section 3).

2 The Nomage corpus and annotation protocol

In this project, we use the French Treebank as our main source of deverbal nouns. In this section, we outline the main features of this electronic corpus, and we describe the deverbal noun candidate extraction process. Then, we proceed by describing our semantic annotation protocol.

2.1 Using the French Treebank for corpus-driven semantics

2.1.1 Overall presentation of the corpus

Abeillé (1999), Abeillé (2001) and Abeillé (2003) describe the French Treebank, a 1 million word electronic corpus for French, following the model of the Penn Treebank (Marcus *et al.*, 1993). The French Treebank is a manually-revised tokenized, lemmatized, tagged and parsed corpus of French-language news extracts taken from the archives of *Le Monde* newspaper articles.

This corpus is, as far as we know, the only manually revised Treebank available for French: other Treebanks have been automatically tagged by members of the VISL consortium (Salmon-Alt *et al.*, 2004), for example, but no corpus of the magnitude and quality of the French Treebank is, as of today, freely available for research purposes. This corpus, similar to the Penn Treebank in its general philosophy, was devised in order to provide linguists with reference data for the manual and automatic construction of formal grammars (HPSG, and other formal frameworks), and to provide the Natural Language Processing community with a benchmark for training and testing automatic taggers and parsers for French.

The French Treebank therefore tries to achieve both coverage [1] and precision: fine-grained distinctions are made, both on the different kinds of tokens which are considered, and on their respective tags (see below). One of the outcomes of this project, in the domain of formal grammars, is the extensive electronic grammar devised by Abeillé and other researchers (Abeillé, 2002). In the domain of corpus-linguistics and NLP, (Kupsc, 2007) is an example of a verbal argument-structure frame lexicon which was semi-automatically derived from the French Treebank syntactic annotations.

In its present version, the corpus comes without any metadata (e.g. date, type of news extract, topic, beginning and end of articles, etc.) other than a unique sentence identifier[2]. The lack of genre, or even author and date metadata, makes the semantic annotation we are conducting non-trivial, as sub-genres such as financial news extracts, for example, generally yield large amounts of hard-to-process deverbal nouns. For example, a very common noun such as *investissement* ('investing') is

[1] Even though the 1 million word standard is now obsolete.
[2] A new version of the corpus is in preparation, though, augmented with these data.

generally used in a deliberately ambiguous manner in financial texts, meaning both the process and result (amount of money) of investing.

Moreover, in the present version of the French Treebank, it is impossible to be sure whether a given sentence where a candidate was found, and the immediately preceding (resp. following) sentence, are part of the same article or not. This means that it is impossible to automatically detect cases where deverbal candidates are used in a generic way (i.e. without their expected arguments) because of an existing anaphoric link with another preceding noun, as is the case of *opération* ('operation'), used in an anaphoric manner to refer to an earlier *bombardement* ('bombing') instance.

2.1.2 Tokens and parts of speech

One of the most distinctive features of the French Treebank with relation to the Penn Treebank is its mixed tokenization: words are tokenized as simple and compound words. This simple/compound distinction is of great importance to the Nomage project, as we explicitly restrict the deverbal noun candidates to be annotated to the set of simple tokens.

The reason for this focus on simple tokens is that, from our experience, the transformations used in our semantic and aspectual annotation protocol generally yield cumbersome results when applied to compound (and more generally multi-word) tokens. Therefore, no clean-cut assessment of their semantic or aspectual properties can be made. More generally, multi-word lexical units, be they true compounds or simple collocations, frozen expressions or even phrasal patterns, are generally not readily available for the transformations we base our annotation protocol on, essentially transformations or rephrasing patterns.

Each sentence in the corpus is associated with an XML tree representing its syntactic structure, which renders the identification of possible PP modifiers of a given noun quite straightforward, for the purpose of describing each noun's syntactic argument structure (and thus its potential semantic argument structure).

Finally, an important feature of this corpus is that only verb *nuclei* adjuncts and arguments are clearly identified, while other predicative words' arguments, such as those of nouns and adjectives, are not. This means that no syntactic argument-structure characterization can be automatically derived from the existing syntactic tags.

2.1.3 Candidate selection policy

For the purposes of our project, the deverbal noun candidates were those simple tokens bearing the "common noun" tag N-C (as opposed to "person name" tag N-P), and bearing one of the following suffixes: -ade, -age, -ance, -ée, -ence, -ment, -sion, -tion, -ure, -xion. An additional constraint was used, on the length of the candidates, in order to filter out false positives, such as *rade*, *page*, *garance*, *rosée*, etc., i.e. candidates which are clearly not derived from any French verb. The total set of candidates to be annotated is 10,584, representing over 110 person/days of annotation, according to the preliminary annotation experiments conducted in 2007 and 2008.

Suffix	Candidates
-ade	24
-age	575
-ance	716
-ée	425
-ence	521
-ment	1824
-sion	1036
-tion	4884
ure	559
-xion	20

Table 1: Absolute frequencies of candidates by suffix.

The Table above gives an overall view of the different suffixes and the respective absolute frequencies for each category of candidates (other peripheral candidates are included, not mentioned above).

2.2 Using rephrasing tests for semantic annotation

One of the main issues our project aims at addressing is to what extent deverbal nouns inherit semantic (in particular, aspectual) features from the verbs they derive from.

The main problem faced in the aspectual study of nominalizations is finding the appropriate linguistic tests, as the habitual ones generally used for verbs (Dowty, 1979) do not apply to nouns. In previous works (Huyghe & Marín, 2007; Haas *et al.*, 2008; Barque *et al.*, 2009) we have proposed several tests that seem to work adequately from a theoretical perspective.

Yet our purpose here is to empirically analyze the aspectual properties of deverbal nouns. To that end, we propose to annotate each deverbal noun from our corpus by using the following ten tests, devised so as to be straightforwardly applied in real contexts:

1. *Plusieurs* N ('several Ns')
2. N *avoir lieu* ('N to take place')
3. *Éprouver/ressentir* N ('to feel N')
4. *Un peu de* N ('a bit of N')
5. N *durer x temps* ('N to last x time')
6. N *se trouver* ('N to be at')
7. *Effectuer/procéder à* N ('to carry out N')
8. *État de* N ('state of N')
9. N *se dérouler* ('N to develop')
10. Cardinal number + N

These tests highlight the main aspectual and referential properties of deverbal nouns. Thus, tests 1, 4 and 10 are aimed at countability; tests 2 and 7 at eventivity; tests 5 and 9 at punctuality; tests 3 and 8 at stativity, and test 6 at result or object readings. Based on the results of each of these tests, we can distinguish four classes of nominals: states (*admiration* 'admiration') durative events (*opération* 'operation'), punctual events (*explosion* 'explosion') and objects (*bâtiment* 'building'), as shown in the following table:

		State	Event		Object
			Durat.	Punct.	
1	Plusieurs	−	+	+	+
2	Avoir lieu	−	+	+	−
3	Éprouver/ressentir	+/−	−	−	−
4	Un peu de	+/−	−	−	−
5	Durer x temps	+/−	+	−	−
6	Se trouver	−	−	−	+
7	Effectuer/procéder	−	+	+	−
8	État de	+/−	−	−	−
9	Se dérouler	−	+	−	−
10	Cardinal	−	+	+	+

Table 2: Description of nominal aspectual classes by means of transformation tests.

As can be observed, there are certain correlations among tests. For instance, a positive outcome for test *avoir lieu* is usually correlated with a positive value for *se dérouler*. The same holds true for *plusieurs* and *cardinal* tests. The rationale behind this apparent redundancy is that we wished to identify the different aspectual properties by using both a generic and a more specific test for each property. This means that a positive outcome for a specific test generally entails a positive outcome for a more generic one. For example, "Avoir lieu" and "Se dérouler" target the same general property, i.e. an event reading, but "Avoir lieu" is more generic, which means events generally accept test 2, and might also accept test 9, but the reverse is not generally true.

As can be seen, our annotation protocol is not a classical one, in which categories are supposedly orthogonal to one another: most of our tests are meant to be correlated, responses to these tests range from "Yes" versus "No", to "Not applicable" and "Don't know". Moreover, even though the "low-level" annotations are already used for lexicographic description validation in our lexicon (see below), the annotation process is not yet complete. Therefore, no inter-annotator agreement metrics can be provided at this date. Nevertheless, we plan to include such metadata in the final version of the Nomage corpus and lexicon.

2.3 Annotating the corpus

In order to keep subjectivity as low as possible during the annotation process, each test must meet certain conditions to be applied. On one hand, tests 1, 4, 8 and 10 could be "directly" applied: *plusieurs* N; *un peu de* N, etc. On the other hand, tests 2, 3, 5, 6, 7 and 9 are applied by means of a relative clause. Tests 2, 3, 5, 7 and 9 follow the pattern N + relative clause + temporal complement, while test 6 follows the pattern N + relative clause + place complement. In these cases, additional constraints are stipulated with respect to: i) the verbal tense of the relative clause; ii) the type and the position of the relative clause; iii) the type of the temporal complement.

To illustrate our annotation protocol, we can take two nominalizations, such as *reconversion* ('career switch') and *rédaction* ('editorial staff'), appearing in two sentences of our corpus, (1), and give responses for three different tests, (2)-(4):

(1) a. Dus à des motifs personnels et à une reconversion dans le commerce de l'art.
'Owing to personal reasons and to a career switch in the art trade.'

b. D'ailleurs, en ce soir de réveillon, la rédaction était réduite à la portion congrue.
'Moreover, this Christmas Eve, the editorial staff was limited to the strict minimum.'

(2) *Test 1: Plusieurs N*
a. Plusieurs reconversions dans le commerce de l'art.
'Several career switches in the art trade.'

b. Plusieurs redactions étaient réduites à la portion congrue.
'Several editorial staffs were limited to the strict minimum.'

(3) *Test 2: N qui avoir lieu + temporal complement*
a. Une reconversion dans le commerce de l'art qui a eu lieu cette année.
'A career switch in the art trade which took place this year.'

b. *La rédaction, qui avait eu lieu la veille, était réduite à la portion congrue.
'The editorial staff, which took place the day before, was limited to the strict minimum.'

(4) *Test 6 : N qui se trouver + place complement*
a. *Une reconversion dans le commerce de l'art qui se trouvait à Paris.
'A career switch in the art trade which was located in Paris.'

b. La rédaction, qui se trouvait à Paris, était réduite à la portion congrue.
'The editorial staff, which was located in Paris, was limited to the strict minimum.'

Therefore, in these contexts, both *reconversion* and *rédaction* should be positively marked with respect to the *plusieurs* test; only *reconversion* passes the *avoir lieu* test, and only *rédaction* passes the *se trouver* test.

		reconversion	*rédaction*
1	Plusieurs	+	+
2	Avoir lieu	+	−
3	Éprouver/ressentir	−	−
4	Un peu de	−	−
5	Durer x temps	+	−
6	Se trouver	−	+
7	Effectuer/procéder	+	−
8	État de	−	−
9	Se dérouler	+	−
10	Cardinal	+	+

Table 3: Overall test annotation for *reconversion* ('career switch') and *rédaction* ('editorial staff').

Table 3 illustrates the behavior of both *reconversion* ('career switch') and *rédaction* ('editorial staff') with respect to the battery of ten tests used in the Nomage annotation task.

3 Building the Nomage lexicon

The Nomage lexicon describes the different nominalizations extracted from the French Treebank, which amount to a total of 815 potential polysemous units. The lexicographic description process is twofold:

- Each unit is associated with a range of semantic properties, based on high-level semantic characterization proposed by the lexicographer;
- The "low-level" annotation data, based on the tests presented above, are used in order to complement the high-level categorizations.

In the second phase of the lexicographic description of each entry, we try to assess whether the high-level word-sense distinctions and semantic properties are correlated with aspectual properties as captured by our annotators.

3.1 Structure and content of the Nomage lexicon

Figure 1 shows the structure and content of the PROMOTION entry of the Nomage lexicon that describes the two sub-entries of PROMOTION that occur in the French Treebank.

Polysemy assessment, and thus word-sense distinctions, are mainly made following a traditional lexicographic approach, by observing each of the 24 occurrences of the noun PROMOTION, together with the habitual semantic descriptions found in a reference dictionary. Each distinguished word-sense or lexical unit (in our case PROMOTION#1 and PROMOTION#2) is associated with the corresponding definition, taken from the *Trésor de la Langue française informatisé* (Dendien & Pierrel, 2003), or a unified definition whenever the word senses in this reference dictionary are too fine-grained. Each lexical unit is associated with an example, taken from our corpus, and a list of pointers to all occurrences in the corpus. Once the nominal unit is identified, we provide information on its source verb, by selecting a matching verbal entry from the syntactic lexicon Dicovalence (van den Eynde & Mertens, 2003). We provide an example to illustrate the sense of each source verb. Furthermore, we complement argument structure information extracted from Dicovalence with an aspectual class tag. In our example, we label the verb PROMOUVOIR#1 as an achievement, and PROMOUVOIR#2 as an activity, based on their respective aspectual behavior.

PROMOTION

PROMOTION#1

Definition: Accession d'une ou plusieurs personnes à un niveau supérieur de responsabilité ou à de meilleures conditions. [An advancement in rank or position.]
Example: *C'est arrivé après sa promotion au poste de directeur financier.* ['It happened after his promotion to a finance director.']
French Treebank occurrences: d1e22886, d1e22934, d1e10709 …
Source verb:

PROMOUVOIR#1
Example: *Ses supérieurs hiérarchiques décident de le promouvoir au poste de responsable d'unité.* ['His superiors decided to promote him to head of division.']
Argument structure: P0 promouvoir P1 (P2)
Aspectual class: achievement

Argument structure: promotion de Y à X accordée par X
Aspectual class: achievement

PROMOTION#2

Definition: Action de provoquer le développement ou le succès de quelque chose. [Cause the development or success of something.]
Example: *Chirac va faire la promotion de son livre en plein marasme judiciaire.* ['Chirac is about to engage in the promotion of his book, while several law suits are being filed against him.']
French Treebank occurrences : d1e71021, d1e10706, d1e44169, d1e63654…
Source verb:

PROMOUVOIR#1
Example : *Le CNRS devait promouvoir la recherche scientifique.*['The CNRS was supposed to foster scientific research.']
Argument structure : P0 promouvoir P1
Aspectual class: activity

Argument structure: promotion de Y par X
Aspectual class: activity

Figure 1: Two entries for the noun PROMOTION in the Nomage lexicon

The verbal valence found in Dicovalence allows the number and the type of arguments of the nominal form to be deduced. Contrary to what can be found in the Nomlex lexicon (Macleod et al., 1998), no detail on the syntactic realization of each argument is provided in our lexicon, for example for PROMOTION#1, we could give the set of possible instances of the syntactic frame description: *sa promotion au poste de directeur financier* ('his being promoted to finance director'), *la promotion qui lui a été accordée par ses supérieurs* ('the promotion his superiors granted him').

Lastly, we provide an aspectual tag for each noun entry. In our example, both PROMOTION#1 and PROMOTION#2 belong to the same classes as their source verbs: i.e. respectively an achievement and an activity.

3.2 Using "low-level" annotations for entry validation

In order to illustrate how we cross-validate the information in our lexicon with collected "low-level" annotations, we will now examine some annotations for noun PROMOTION. We illustrate below 4 of the 24 sentences in which noun PROMOTION occurs.

(5) Les moyens à la disposition des opérateurs publics concourant à la **promotion** des ventes françaises au Japon augmenteront de plus de 40%.
'The financial incentives available to public-owned companies that actively support French business transactions in Japan will increase by over 40%.'

(6) L'infatigable patron de Lancôme (groupe L'Oréal) en Allemagne ne ménage pas son temps pour la **promotion** de son entreprise.
'The tireless chief executive of Lancôme's (l'Oréal group) German division spares no efforts in promoting his company.'

(7) C'est arrivé après sa **promotion** au poste de directeur financier.
'It happened after his promotion to finance manager.'

(8) La première **promotion** est sortie en 1991, à notre grande satisfaction.
'The first class completed their program in 1991, to our great satisfaction.'

Sentences (5) and (6) exhibit the same outcomes for each test, which indicates two occurrences of the same lexical unit. The fact that the tests registering positive outcomes (tests 5 and 9) are those associated with the durative aspect shows that we are facing occurrences of the unit PROMOTION#2 (cf. entries in Table 4).

Annotations for sentence (7) show a whole different picture: a positive outcome for test 2 allows us to categorize this lexical unit as an event, but since tests 5 and 9 (related to duration) both register a negative outcome, it can be asserted with some degree of confidence that this occurrence does not have a durative reading in that sentence. Therefore, it appears we are faced with the unit PROMOTION#1, which has an achievement reading.

Lastly, test outcomes for sentence (8) hint at the existence of a new lexical unit PROMOTION#3, denoting "a group of candidates or students belonging to the same year level of a given class" (i.e. PROMOTION "school (or other educational institution) class" as in "class of 1980"). The only positive tests for this occurrence are associated with countable (tests 1 and 10) and concrete nouns (test 6).

		Sentence			
		(5)	(6)	(7)	(8)
1	Plusieurs	−	−	−	+
2	Avoir lieu	−	−	+	−
3	Éprouver/ressentir	−	−	−	−
4	Un peu de	−	−	−	−
5	Durer x temps	+	+	−	−
6	Se trouver	−	−	−	+
7	Effectuer / procéder	+	+	−	−
8	État de	−	−	−	−
9	Se dérouler	+	+	−	−
10	Cardinal	−	−	−	+

Table 4: Aspectual test outcomes for 4 occurrences of the noun PROMOTION

These examples show how the collected "low-level" annotations are used in the high-level lexicographic description of each lexical unit of the Nomage lexicon.

4 Further research

We have presented the Nomage project, an ongoing corpus-based semantic annotation project. The resulting annotated corpus forms the ground work for one of the main outcomes of our project: a semantic electronic lexicon for French deverbal nouns. This lexicon will be the first, so far as we know, to propose a description of aspectual properties for French nouns, in the continuity of projects such as Nomlex (Macleod *et al.*, 1998) and SIMPLE (Bel *et al.*, 2000).

Upon completion, the Nomage project will therefore yield semantic annotations for deverbal nouns, based on the French Treebank – as far as we kknow, the only available Treebank for French – together with an electronic lexicon for French deverbal nouns, and a corpus-based semantic annotation protocol.

In future work, we intend to extend the semantic annotation process to French deadjectival nouns (e.g.: *fidelité*, from *fidèle*), and to non deverbal predicative nouns (e.g.: *crime*, *meurtre*). We also intend to extend our semantic annotation protocol to other languages: Spanish, English and Catalan.

References

Abeillé, A. (ed.) (2003). *Treebanks, Building and Using Parsed Corpora*. Kluwer Academic Publishers, Dordrecht/Boston/London.

Abeillé, A. (1999). A Tagged Reference Corpus for French. In *LINC'99 Proceedings*, EACL.

Abeillé, A. (2001). The Paris 7 Annotated Corpus for French: some Experimental Results. In Wilson (ed.), *Corpus Linguistics*, Lancaster.

Abeillé, A. (2002). *Une Grammaire Électronique du Français*, CNRS Editions, Paris.

Barque, L., R. Huyghe, A. Jugnet, R. Marín (2009). Two types of deverbal activity nouns in French. *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pise, 17-19 September, pp. 169--175.

Bel N., Busa F., Calzolari N., Gola E., Lenci A., Monachini M., Ogonowski A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons, *Proceedings of LREC 2000*, Athens.

Chomsky, N. (1970). Remarks on Nominalizations. In A. J. Roderick & P. S. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Waltham, Ginn.

Dendien, J. & Pierrel, J.M. (2003). Le Trésor de la Langue Francaise informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues (T.A.L.)*, 44(2), pp. 11--37.

Eynde, K. van den & Mertens, P. (2003). La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies* 13, pp. 63--104.

Grimshaw, J. (1990) *Argument Structure*. Cambridge: MIT Press.

Haas, P., R. Huyghe, R. Marín (2008). Du verbe au nom : calques et décalages aspectuels. In J. Durand, B. Habert & B. Laks (eds.), *Congrès Mondial de Linguistique Française*, Paris : Institut de Linguistique Française, pp. 2051--2065.

Huyghe, R & R. Marín (2007). L'héritage aspectuel des noms déverbaux en français et en espagnol. *Faits de Langues*, vol. 30, pp. 265--274.

Kupsc, A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Actes de la conférence TALN 2007*, ATALA.

Lees, R. (1960). *The Grammar of English Nominalizations*. The Hague, Mouton.

Ludovic, T. & Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web, *Actes de la 9ième Conférence Annuelle TALN* Nancy.

Macleod C., Grishman R., Meyers A., Barrett L., Reeves R. (1998). NOMLEX: A Lexicon of Nominalizations, *Proceedings of EURALEX'98*, Liege, Belgium.

Mitchell P. M., B. Santorini, M.A. Marcinkiewicz. (1993). Building a Large Annotated Corpus: the Penn Treebank, *Computational Linguistics*, vol. 19, 2, pp. 313-330.